Computer Assessment System for English Communication

Rieko Matsuoka

National College of Nursing, Japan

These days, computers have been utilized in various phases of educational life. As Chappelle mentions (2001, p. 95), computer-assisted testing started to affect more people. In our institution, computer-based system for assessing English proficiency called CASEC was introduced in 2002. Students take advantage of this system which allows them to take a computer-based test as much s they like for a fixed rate fee without any additional cost. In this paper, I would like to explicate this computerized testing system and to indicate its reliability and validity in order to find out if this system is useful as a measuring tool for English learners. Reference is from homepage, handouts, direct communication with a senior researcher who developed this system and wrote a paper on this assessment system.

What is CASEC?

CASEC is the Computerized Assessment System for English Communication originally developed by the Society for Teaching English Proficiency, Inc (STEP, hereafter). STEP, which is the largest testing institution in Japan, spent more than seven years for fundamental research of its development. At present, The Japan Institute for Research on Testing, Inc. (JIEM, hereafter) took over the operation and research on CASEC. JIEM was established in 2000 as an Obunsha Group company and has taken responsibilities of activities related to the research and development of accurate measuring techniques or testing in the field of education (http://casec.evidus).

Individual English communication ability is assessed by utilizing a system of computerized adaptive testing (CAT hereafter). Item response theory (IRT hereafter), which is one of the latest

testing theory, enables this system to assess an individual's proficiency accurately and objectively, according to JIEM (2002, 2004). Specifically, the difficulty level changes, depending on whether a test taker answers the previous question correctly or incorrectly. As a result, the test takers do not have to take questions that would be too easy or too difficult for them. This testing system also enables the testing time to be shorter than other standardized test such as TOEIC that requires more than two hours. The testing time of this computerized test ranges from 20 minutes to 70 minutes with the average time of 40 minutes, based on the report by JIEM (2002, 2004). Additionally with this absolute evaluation system, the CASEC test can evaluate the test takers relatively as well. The details of this system will be explained in the following section.

Contents of CASEC test

There are four sections in this test system. Each of them as follows.

Section 1

Section 1 is for vocabulary knowledge. There are fifteen questions and asking form is four-option multiple choice. Total point is 250 and time limit is 60 seconds per question.

Section 2

Section 2 is for checking the knowledge of idioms and useful expressions. There are also fifteen questions in the form of four-option multiple choices. Total point is 250 that is also the same as Section 1, but time limit is 90 seconds per question, which is 30 seconds longer than the one in Section 1.

Section 3

Section 3 is a listening test, with the purpose of grasping the main point. There are also fifteen questions with four-option multiple choices. The total point is also 250, and the time limit is 60 seconds, which is the same as a vocabulary section (Section 1).

Section 4

Section 4 is a dictation test consisting of ten question. The test takers input the words using

keyboard. Total points are also 250, and time limit per question is 120 seconds, which is the longest of all sections.

Since the individual test taker processes the test, the time spent is varied from 20 minutes to 70 minutes from the beginning of test to the indication of the test score, followed by the end of test. The approximate average time is 40 minutes according to the statistics that JIEM (2004) took.

The rationale of these sections

According to the article from the institute operating CASEC (JIEM, 2004), the contents of CASEC test are rationalized to gain validity as a system of testing English communicative proficiency via internet.

The following is based on the handouts (Hayashi, 2004) and personal communication (Hayashi, 2004).

First, the four sections were selected based on the following two criteria.

(a) The degree of measuring the speaking ability or communicative competence.

(b) The questions which can be applicable to the computer based testing or testing form which does not need to be altered from the present pencil-and-paper test.

Examining dozens of testing forms, the institution selected 18 forms. The research staff at STEP prepared two sample tests (Sample test A and B) using the selected 18 forms, such as vocabulary, conversation, or dictation. Then Sample test A composed of 100 items and Sample test B composed of 72 items were administered to 683 participants. The participants were also given an interview test with a native speaker of English for approximately 20 minutes for each of them. In order to have consistency with the outside criteria, they selected the participants with an experience of taking STEP test – both pencil and paper test, and interview test, which results were known to the research staff at the Institute.

Analysis

Analysis was performed in order to find out the appropriate forms of testing to predict the speaking ability the best. Therefore, the intra-correlation of sample test with interview test and interview test of STEP was extracted. The exploratory factor analysis was conducted in order to see the latent factors as well. As a result four forms were selected for CASEC. Some forms were eliminated because of the difficulty in administering using computer website.

A form for measuring vocabulary did not have a strong correlation with interview tests; however, it was selected because vocabulary was the most heavily loaded factor in English proficiency. The form for vocabulary measurement became Section 1 in CASEC.

A form regarded as measuring expressing ability gained a high correlation with interview test. Consequently, this form was regarded as a high predictor of speaking ability, and became Section 2 in CASEC.

They decided that a listening section should be included. There were several different forms of listening forms, and the one with the highest correlation with the results of interview test was selected as Section 3 of CASEC. Three forms from other listening test forms were integrated into the dictation form, which became Section 4 in CASEC.

Dictation form has the highest correlation coefficient, which was $r = .701$. This means that the dictation form testing can explain nearly 50 % of the interview testing, which requires speaking ability. However, in reality, this statistical data may hold true only for Japanese participants who have studied mainly in Japan. Some of my students who stayed in an English-speaking country for a certain period of time did not gain a high score in Section 4, though they are good speakers of English. This fact may lead to the assumption that dictation can predict the speaking ability of Japanese learners of English given that they are unable to have a chance to expose themselves to an English speaking community. Specifically, if they learn English where it is spoken, they do not need to spell the words in an accurate way in order to speak English. In other words, all good speakers of

English do not necessarily spell the words correctly. However, in the case of learners destitute of sufficient amount of exposure to a native speaking environment, dictation can predict the speaking ability, because they can speak good English when they study English very carefully and are good at spelling words correctly. Actually, in a personal communication with a person in charge of developing CASEC and performing these statistical analyses (Hayashi, June, 2004). it was found that almost all the participants learned English mainly in Japan

CAT

Testing systems using computers are generally called Computer-based Testing (CBT hereafter) (Chappele, 2001) and CASEC is regarded as one of them. Furthermore, CBT has a subcategory called 'Computerized Adaptive Testing' (CAT hereafter) (Chappele, 2001), and CASEC is classified into the subcategory of CAT. As the basic concept of CAT, computers assume the proficiency level of test takers at each item in the process of testing, and choose the best item for each test taker from the item bank that stores sufficient number of items This system promises the accuracy of measurement and smaller number of items can measure the proficiency level of test takers. When a learner takes CASEC for the first time, he or she will get an item of mid level. If the learner can answer it correctly, a harder item will be provided. In the case the learner cannot answer it correctly, an easier item will be provided. The computer judges the level of test takers each time and by so doing test takers do not need to take too easy or too difficult questions. Either too easy or too difficult question fail to measure the proficiency level. The main purpose of CAT is to eliminate the useless process of measurement.

Computers enabled the system to be realized because computers calculate the assumed proficiency level of test takers and select the appropriate item immediately, based on the patterns of response and the value of a given item (Hayashi, 2001)..

IRT

The following explanation about Item Response Theory (IRT, hereafter) is mainly based on

Hayashi (2001).

CAT requires item parameter, which can be compared with each other using the same scale in storage of items. The preliminary study for CASEC had about 3000 question items and all of them were previously tested with 2000 trials. They were all given item parameters using maximum likelihood estimates based on three parameter logistic model. Ten pilot studies were conducted by STEP research group in order to gain item parameters and each study was conducted towards the different subjects; however, repeating the equating work, all items can be compared with on the common scaling system. Maximum likelihood of estimation of proficiency level of each subject was able to be measured on the same scaling system. Three parameter logistics model is the formula that calculates the probability of answering a certain item question (*j*) by a test taker whose proficiency level is θ . The item is given by three parameters *aj (discrimination power), bj(difficulty level) and cj(pseudo-chance level).*

The formula is:

$$Pj(\theta) = cj + (1-cj)/1 + exp[-Daj(\theta - bj)]$$

D in the above formula, which is a constant, is 1.7 in order to make the logistic line closer to cumulative normal distribution. The maximum likelihood estimation here is to find θ which makes the probability of emergence the maximum under the ansering pattern given to each item by a test taker in the list of items with known item parameters.

Define $uj = 1$, when a test taker gives a correct number to an item j and, $uj = 0$, otherwise. The probability of getting an answering pattern, which is $u = [u1, u2, \ldots uj \ldots, un]$, in the *n* items by a test taker, is gained by the following formula on the condition that each item is independent. This is called an assumption of local independence.

$$L(\theta|u) = \text{II}jPj(\theta)\, uj\, Q(\theta)1\text{-}uj$$

This is a likelihood function in an answering pattern given to a person with proficiency value 0. In that case, $Q(\theta) = 1\text{-}Pj(\theta)$. In the above formula, the value of θ is unknown and the most likely

value of θ is explored within the possible range, and that value is regarded as the estimated proficiency level of a test taker to indicate the given response. This process is called maximum likelihood estimation. The value of θ is calculated by logarithm of the above formula and differentiated θ should be zero. This process is repeated until the gap between $\theta r+1$ and θr become smaller than the previously measured value, which means the estimation is precise enough.

Verification

Validity was verified by comparing CAT with pencil & Paper test (PPT hereafter) (Hayashi, 2004a). One hundred sixty-eight monitors took both CAT and PPT, and the CAT gave them different items depending on their results of the previous item; however, the PPT, which is a traditional form of testing, gave the monitors the same exam. Among them, 48 monitors took the CAT more than once. Both reliability and validity were gained based on these data, according to Hayashi (2004a, 2004b).

The following results are based on Hayashi (2004a, 2004b).

Comparison of CAT and PPT

Maximum likelihood estimation was conducted for both CAT and PPT. Correlation coefficients for the score values computed of CAT and PPT for each section were .865 to .899. Correlation coefficient for total test was .958. Standard Error of measurement for CAT is smaller than that of PPT, despite CAT has fewer items. The smaller standard error of measurement is one of the strengths that CATs have. The study at the institution (JIEM) indicates that the item numbers of CAT can be reduced by 40 percent from PPT. This means that CAT is a more capable instrument to measure English proficiency than PPT, according to Hayashi (2001).

Reliability

In the experiment at JIEM, the same participants conducted CAT several times in order to examining the reliability of CAT, using parallel test method. Forty-eight participants took CAT three times. Because the CAT (CASEC) is based on the item response theory, θ (inferred proficiency

value) was compared by the same measurement. Average value and standard deviation of θ, correlation coefficients of θ from the handout (Hayashi, 2004) are shown below.

Inferred proficiency value (θ)

	Average value	Standard deviation
1.	423.2	42.6
2	*425.2*	42.5
3	423.7	42.9

Correlation coefficients

	1	2	3
1.	1.000		
2.	.975	1.000	
3	.968	.964	1.000

As shown above, high reliability coefficients were gained. According to Hayashi (2001), these coefficients were found to be higher than those of PPT.

Comparison to standardized tests

Based on the results of approximately 2,000 examinees, the CASEC test is correlated to the TOEIC, TOEFL and STEP test. The correlation coefficient with TOEIC was found out to be .86 (Maruzen, 2003b).

Concluding remark

The Japan Institute for English Measurement (JIEM) verified both validity and reliability, and indicated the high correlation coefficient with TOEIC test. JIEM, according to Hayashi (2004a), administered a questionnaire to the participants about their impression on CASEC, and gained favorable responses. As Chappele (2001) points out, test method influences test takers' performance, and in Computer Assisted Language Tests, the students who are resistant in using computers, may produce negative results. However, in my institution, most students have enjoyed taking the CASEC test. Though taking the CASEC test is not mandatory for them, some students have taken it regularly.

The CASEC test, which is easily accessible to learners, seems to be a promising tool to measure the English communicative proficiency. According to the periodical report issued by an agency (Maruzen, 2003a), the number of test takers has been increasing since they started the CASEC service in 2002. As of July, 2004, approximately one hundred twenty thousand people including college students and company employees are taking the CASEC test in total. Accordingly, the times of questions stored for the assessment system also increased from 2000 in 2002 to more than 4000 in 2004. The researchers at JIEM seem to keep on improving the quality of this system, and utilizing such a computer system at an institution will enhance the quality of language learning if it can be used properly.

References

Chappelle C. A. (2001). *Computer applications in second language acquisition*. Cambridge: Cambridge University Press.

Hayashi, N. (2001). EIGO NORYOKUSOKUTEINIOKERU CAT NO TEKIOREI TO KOKASOKUTEI [Case study of CAT and its effect measurement in English proficiency assessment]. *KEISOKU TO SEIGYO [Measurement and control]*. *4*(9). 572-575.

Hayashi, N. (June, 2004). Personal communication.

Hayashi, N. (2004). KEISHIKIKETTEI SAMPLE TEST MONDAIREI [Sample test for deciding the testing formats].

Japan Institute for Educational Measurement (JIEM). (2002). From selection to placement. *Kyoikusokutei-kenkyusho*. 1-3.

Japan Institute for Educational Measurement (JIEM). (2004). The computerized assessment system for English communication (CASEC). http://casec/evidus. 1-2.

Maruzen. (2003). INTERNET O RIYOSHITA KOJINTEKIOGATA COMMUNICATION NORYOKU-TEST NO GOTEIAN [Proposal of computerized assessment system for English communication]. MARUZEN, 1-14.

Maruzen. (2003). TATEST TONO SOKAN/BUNPU SHIRYO [Correlation with other test and scatter plot]. MARUZEN, 1-3.